AN OVERVIEW OF RESULTS FROM RUTGERS’ INVESTIGATIONS OF INTERACTIVE INFORMATION RETRIEVAL*

Nicholas J. Belkin

ABSTRACT

Over the last four years, the Information Interaction Laboratory at Rutgers’ School of Communication, Information and Library Studies has performed a series of investigations concerned with various aspects of people’s interactions with advanced information retrieval (IR) systems. We have been especially concerned with understanding not just what people do, and why, and with what effect, but also with what they would like to do, and how they attempt to accomplish it, and with what difficulties. These investigations have led to some quite interesting conclusions about the nature and structure of people’s interactions with information, about support for cooperative human-computer interaction in query reformulation, and about the value of visualization of search results for supporting various forms of interaction with information. In this discussion, I give an overview of the research program and its projects, present representative results from the projects, and discuss some implications of these results for support of subject searching in information retrieval systems.

INTRODUCTION

Researchers associated with the Information Interaction Laboratory at Rutgers’ School of Communication, Information and Library Studies have been engaged in a program of research aimed at understanding and supporting the information-seeking behaviors of people in advanced interactive information retrieval (IR) systems. This program had its roots in

*Research supported in part by a DARPA Equipment Grant, by NIST Cooperative Agreement No. 70NANB5H0050, and by TIPSTER Phase III Contract No. MDA904-96-C-7297
several earlier research projects. The first of these was a study led by N. J. Belkin and T. Saracevic on the design of third-generation Online Public Access Catalogs (OPACs) (Belkin et al., 1990). This study intended to identify design principles for third-generation OPACs through an understanding of the intentions and behaviors of people as they interacted with information in a variety of libraries. The results of this study included the development of a methodology for investigating goals, behaviors, and intentions associated with interaction with information; classifications of behaviors and intentions in libraries; identification of some relationships among goals, behaviors, and intentions; and the beginnings of identification and classification of a variety of so-called “information-seeking strategies.”

The second “precursor” project was conducted in collaboration with the Information Retrieval Service of the European Space Agency (ESA/IRS). This project, as reported in Belkin and Marchetti (1990); Marchetti and Belkin (1992); and Belkin, Marchetti, and Cool (1993), was concerned with the development of an interface to an existing commercial IR system that would support seamless movement from one kind of information seeking strategy (ISS) to another—e.g., from browsing through a thesaurus to searching on a specified topic. This study combined the methods of cognitive task analysis with support for multiple ISSs and introduced a classification of ISSs based on the four dimensions of user’s goal, method of searching, mode of retrieval, and type of information interacted with. This work was based on previous research by both us and others on information-seeking behavior, but its design was somewhat limited by the necessity of implementation within an existing system framework. The results of this study included the identification and classification of multiple ISSs and the design and construction of an interface based explicitly on an analysis of information-seeking behavior that supported multiple ISSs.

The third precursor project was conducted in collaboration with colleagues at the Gesellschaft für Mathematischen Datenverarbeitung Institute for Integrated Publication and Information Systems (GMD/IPS) (Belkin et al., 1995). In this work, we built on our earlier studies described above and combined these ideas with two other theoretical stances—dialogue-based interaction and case-based reasoning. The idea here was to develop a new system that would support user interaction with information through a set of structured dialogues, each of which was specific to a particular ISS in its general structure, and which would be linked to one another in order to support changes from one ISS to another according to generalizations built from a library of cases of information-seeking episodes. This project resulted in a prototype system, MERIT, which supported several different ISSs and structured changes among them in a mixed-initiative dialogue.
The common thread among these three projects was the attempt to understand and characterize the variety of people's interactions with information in a way that would be useful for the design of systems to support such interactions. Perhaps the most important results of these projects were the development of methods for the study of interactions with information and for the identification of the goals and intentions leading to different interactions; the initial attempts at characterizing ISSs according to a set of behavioral dimensions or facets (see Figure 1); and the establishment of principles for drawing relationships between different ISSs and IR design features for supporting them. In particular, the dimensionalizing of ISSs led to the concept of a person moving about in a space of possible ISSs according to both planned and situated action.

The results of the three projects described earlier have provided a basis for a long-term research program at Rutgers on the relationships between people's interactions with information (particularly, although not exclusively, their information-seeking behaviors); the goals associated with different interactions; and design specifications for IR systems. The goals of this research program are to develop highly interactive IR systems which are designed to respond to people's characteristics, goals, intentions, and behaviors, in particular by appropriate and effective support of their interactions with information. Although much of this program has been carried out within the framework of our participation in the Text Retrieval Conference (TREC) program's Interactive Track (see, for example, Harman, 1996), it has also included Ph.D. dissertations based on these data and projects supported by the National Institute for Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). In the following sections, I briefly describe a number of the specific projects that have been carried out in the course of this program and summarize some of their more important results.

GENERAL STRUCTURE OF THE RESEARCH PROGRAM

As mentioned earlier, we have been primarily working within, and extending, the TREC Interactive Track program (for a detailed description of TREC, see, for example, Harman, 1996). Here, I give a brief summary of the general characteristics of this program to provide a context in which to interpret the description of our specific projects.

TREC is a program for the comparative evaluation of IR systems and their components. It is organized by NIST and is based on the DARPA-funded TIPSTER (not an acronym) program of research in information retrieval and information extraction. One product of the TIPSTER program was a so-called IR test collection. This provided the impetus for the TREC program, which is based on a (growing) test environment for IR systems, consisting of: a database of the full text of documents from a
**DIMENSIONS**

<table>
<thead>
<tr>
<th>ISSs</th>
<th>Method</th>
<th>Goal</th>
<th>Mode</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sc</td>
<td>S</td>
<td>L</td>
<td>S</td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>8</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>11</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>12</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>13</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>14</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>16</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

**Method:** Sc=Scan; S=Search

**Goal:** L=Learn; S=Select

**Mode:** R=Recognize; S=Specify

**Resource:** I=Information; M=Meta-information

Figure 1. Dimensions and Types of Information Seeking Strategies (ISSs) (Belkin et al., 1993, p. 326)

A variety of information sources—e.g., *The Federal Register*, the *Wall Street Journal*, the AP Newswire, US Patents, selected Ziff-Davis publications, the *San Jose Mercury-News*, the *Financial Times*, the *Congressional Record*, and some sources in Japanese, Spanish, and Chinese—which now amounts in total to over 1 million documents and approximately 4GB of data; a set of topic or information problem descriptions growing at the rate of 50 per year that at the moment total 300; and relevance judgments for as many as 3,000 documents for each of the topics.

The general structure of TREC is that groups that wish to participate in the program are provided the test collection and are set several IR tasks that must be completed by all participants under a set of quite strict rules and under some specific time constraints. Having performed these tasks and submitted their results to NIST, the participants in each year’s TREC meet at a small conference to present their results, to discuss them, to compare them with others, and to try to understand how various techniques work in improving retrieval performance (or not) and why. TREC-1 (the conference) was held in autumn 1992 (the work reported at the conference was carried out during the preceding nine months); the pro-
gram is now up to TREC-6 and has about sixty-five participating groups from all parts of the world.

In the past several years, this general structure of TREC has evolved from requiring all participants to complete the same two tasks to the establishment of a number of different "tracks," each of which is concerned with some particular task or problem in IR. One such track is the "Interactive Track." The purpose of this track is to investigate the performance of interactive IR systems—i.e., systems in which the user of the system plays an active role. This poses especially difficult evaluation problems; so much of the work of the Interactive Track has been in trying to develop methods for characterizing, evaluating, and comparing interactive IR systems. Groups at Rutgers have participated in the Interactive Track at TREC conferences 3 to 6 and, in the project descriptions below, the various evaluation measures and IR tasks that have been used at different times will be described in more detail.

In the studies that we have done investigating interactive IR, we have used the InQuery retrieval engine from the University of Massachusetts (Callan et al., 1992) as our basic IR system, modifying and controlling various elements for different studies. InQuery is a best-match IR system based on a probabilistic inference network formalism that allows both structured and unstructured queries, does sophisticated automatic indexing, and supports a variety of retrieval features including relevance feedback. Its richness of capabilities and its flexibility have allowed us to conduct a variety of studies of information-seeking behavior in widely different contexts.

The general progression of these studies has been from initial observational and descriptive studies of information-seeking behavior, to experimental and analytic studies in which we attempt to test the effect and effectiveness of various features and conditions. Each study follows the general pattern of having volunteer subjects participate in an experimental session in our Information Interaction Laboratory. These sessions begin with the administration of a background questionnaire in which demographic data and data about the subject's experience with IR systems of various types are elicited. This questionnaire is followed by an entry interview, in which data about the subject's previous experience, and other characteristics relevant to the goals of the specific study, are elicited and tape recorded. On completion of the interview, the subject is administered a hands-on interactive tutorial in the particular IR system in which the searching will be done. This tutorial ranges in length from twenty-five to forty-five minutes and either includes a practice search on the system for the task that the subject will be asked to perform or uses an example task within the tutorial itself. After the tutorial, the subject is introduced to the searching task which he or she will be asked to perform. These tasks are all based on TREC topics but vary from study to study according to the TREC Interactive Track task that is set that year or according to
the experiment which we are running. The specific task is described in detail on a paper form that the subject is given at this time. The subject then performs the given IR task(s) and is instructed to “think aloud” during the task performance. The entire search is automatically logged, and the thinking aloud is recorded on videotape along with a view of the monitor. After the performance of each IR task—i.e., each individual search—the subject is asked to fill out a “search evaluation form,” which is a scaled self-report on the subject’s knowledge of the topic and of various aspects of the subject’s perception of difficulty of the task and success in the task. Each subject engages in at least two such searches (the number depends on the specific study). After the assigned tasks are completed, the subject is administered an exit interview, which elicits information about the subject’s searching experience that is relevant to the specific study. This interview is tape recorded. The entire experimental session ranges in length from two to three hours.

The questionnaire responses, the tape-recorded interviews, the thinking-aloud protocols, the search logs, and the subject’s performance in the task (measured by the relevant TREC measures) are thus the basic data which we analyze in a variety of ways in order to understand and characterize the nature of the person’s information-seeking behavior and interaction within the IR system. Methods of analysis include content analysis, sequential analysis, standard descriptive statistics on the questionnaire data and on the log data, and inferential statistics relating various characteristics of the interaction to one another and to performance.

DESCRIPTIONS OF STUDIES IN THE RESEARCH PROGRAM

TREC-2: Combining Evidence for Information Retrieval

The first study that we conducted in the TREC research program was concerned not with interactive IR but rather with the issue of the effect of combining different query representations of the same information problem on IR performance. The results of this study, and a closely related pre-cursor, have been reported in Belkin, Cool, Croft, and Callan (1993); Belkin, Kantor, Cool, and Quatrain (1994); and Belkin, Kantor, Fox, and Shaw (1995).

The reason for this study was to follow up on previous research (e.g., Saracevic & Kantor, 1988; Turtle & Croft, 1991) which had indicated that the combination of different representations (i.e., different queries) of the same information problem would lead to better IR performance than that achieved by any one of the representations alone. To investigate this issue, we designed a non-interactive study in which seventy-five volunteer expert searchers (that is, professional information scientists/librarians who had been engaged in online searching for one or more years) were each
each asked to construct Boolean queries for five different TREC topics. There were seventy-five such topics consisting of a title, a one- or two-sentence description of the topic, a two- or three-sentence narrative which defined conditions of relevance, and sometimes concepts and definitions associated with the topic. These topic descriptions were sent to the volunteers, who constructed the queries in their own time (but informing us of how much time they spent on each), according to the query language with which they were most familiar, and returned them to us by mail. The resulting queries (five different ones for each of the seventy-five topics) were translated by the experimenters into the InQuery retrieval language, and then a series of experiments were run against the TREC databases with different levels of combination of the queries.

The results of this study showed that combining these query representations within the InQuery retrieval model led to increasing average performance according to the number of queries that were combined—i.e., performance was best overall when all five queries for each topic were combined into a single query. This result has interesting implications for interactive IR, since it suggests that if IR systems encourage people to represent their information problems in alternative ways, performance will increase; it also suggests that if IR systems are designed to automatically produce alternative query representations, performance will increase. An interesting result of a small experiment conducted during this study was that, if all of the operators (e.g., AND, OR, NOT, proximity) were removed from the combined queries—i.e., if the queries were changed from structured to unstructured or natural language—performance was somewhat better than that for the structured queries. This suggests that, at least for the best-match retrieval model underlying the InQuery system, elaborately structured queries are not necessary to get good IR results.

**TREC-3: New Tools and Old Habits**

TREC-3 marked the first truly interactive IR study we carried out in this program. The results of this study are reported in Koenemann, Quatrain, Cool, and Belkin (1995). The problem that we addressed in this study was to learn how people who were already expert in searching in conventional IR systems—i.e., exact-match systems—would understand, adopt, and adapt to IR systems which had quite different features. The rationale behind the study was that quite soon, operational systems based on new retrieval models, and incorporating such features as ranked output, relevance feedback, and unstructured query input, would become quite common, but that very little was known about how people would understand and react to such features. Our assumption was that, in order to implement these kinds of features in the most effective ways, one would need such knowledge.

In this study, we asked ten expert searchers (the sample population and expertise were defined as in our TREC-2 study) to perform five
different searches on our experimental system using the TREC routing task (this involved searching for documents on a specified topic in a particular database and, on the basis of the search, to construct a query which would be used as a routing or filtering query against a new database). This study was observational and descriptive in nature, and there was no attempt to control by, for instance, having each subject search on the same five topics. Rather, we grouped the fifty TREC topics into ten groups, each of which we tried to make as heterogeneous as possible in terms of domain and difficulty, and thus as much alike one another as possible, and assigned each group to one subject. The version of InQuery that we used in this study allowed both structured and unstructured queries (and combinations of both), gave relevance-ranked output, and supported automatic relevance feedback. The database was, of course, full text, and the only form of indexing was statistically-based automatic indexing—i.e., no controlled vocabulary indexing. Because of the flexibility in the query language, searchers could choose to use their familiar structured-query methods or to use unstructured query input and the new features which supported query modification or any combination thereof.

There were a number of interesting results which came from this study, including the fact that relevance feedback was used by almost every subject at least once, and that people were, overall, fairly effective in the new environment. But perhaps the most interesting results concerned the relationships between what we interpreted as the nature of the subjects' models of IR, their adoption or incorporation of new IR tools, and their performance in the task. In examining the subjects' searching behaviors, we found that some people, who had a great deal of experience in conventional IR and who adapted the features of our version of InQuery to support their normal searching behaviors, had quite high performance. But some subjects, who had relatively little IR experience, and who adopted the new features to a large extent in new searching behaviors, also had high performance. But some other subjects, at a variety of levels of experience, charted an intermediate course, attempting to adapt aspects of the new features to support their normal searching behaviors; these people did relatively poorly in the task. These data led us to speculate that, although people could indeed use the new features effectively with relatively little training, their adoption of the new features, and their success in interaction, depended in some way on the nature and strength of their mental model of IR. These results led us directly to the study we did for TREC-4 and also to an experimental investigation of various implementations of relevance feedback, both of which were funded by NIST.

**TREC-4: Relevance Feedback and Ranking in Interactive IR**

Our participation in TREC-4 (Belkin et al., 1996; Cool et al., 1996) was concerned with understanding how people's mental models of IR af-
fect their behaviors in new (to them) IR systems—i.e., we were trying to follow up on some implications of our TREC-3 study. In order to do this, we designed the study to obtain descriptions of people's "normal" searching behaviors (this would give us a representation of their mental models of IR), and to give them a system with only a minimal number of new features that would not support normal exact-match searching behaviors. The version of InQuery we used offered only unstructured query input, and its primary features were relevance-ranked output of titles, display of the full text of any specified retrieved item, automatic relevance feedback, and the ability to save documents. We recruited fifty volunteer subjects from a rather different population than previously in order to have a range of experience with IR systems represented.

The task the subjects were given in this study was the TREC ad hoc task of retrieval of as many good documents as they could find within thirty minutes. Each subject did a practice search after the tutorial and two experimental searches. There were twenty-five topics in all, and they were distributed among the subjects so that there were four searches for each topic, each by a different subject. We could thus compare performance between subjects on the same topics.

In order to elicit data that could give us a way to characterize the subjects' mental models of IR, in the entrance interview, subjects were presented with an example information problem (a TREC topic of the sort that they would encounter in the experiment), and asked them: (1) how they would go about planning a search on this topic in any IR system that they were familiar with; (2) what their initial query would be; (3) how they would decide if they had retrieved any good documents; (4) what they would do if there were no good documents retrieved; and (5) how they would know they were finished with the task.

Again, there were a number of interesting results from this study, including several unexpected ones. As a byproduct of our entrance interview, we developed a classification of "normal" information-seeking strategies in interactive IR, which both confirmed some previous work (especially Bates, 1979) and extended it. In particular, we found evidence for previously unreported strategies based explicitly upon interaction with the database as a whole, with retrieval results, and with the individual information objects. With respect to the mental model question, our preliminary findings are that the confidence in, or strength of, the model is more predictive of successful performance than is degree of experience with IR systems, which was not related to performance. Also, it was found that there were several patterns of adoption/adaptation that were related to strength of mental model. But perhaps the most interesting finding of this study was that, although subjects uniformly found ranked output useful and although all subjects but two used relevance feedback—and all of those found some utility in relevance feedback—there were some
consistent problems that people had in using these features. Most of these could be considered as *communicative* problems in that the concerns that were expressed had to do with the inability of the subject to communicate her/his intentions and problems accurately to the rest of the system ("the system doesn't seem to understand me"), with the system's inability to communicate its operations and behavior to the subject ("I don't understand why it's doing this"), and with the lack of cooperation between subject and system ("I wish I could tell it what I mean or influence its behavior more"). This result has led us to consider that, to make best-match, relevance feedback-based IR systems effective and usable, it will be necessary to design them in ways which give the users in such systems more control over various operations, and which display, in more obvious ways than lists of titles, why it is that the system obtains the results that it does, and why it has operated in the way that it has. One specific implication of this work on system design is that negative relevance feedback could be important in interactive IR (Cool et al., 1996).

**NIST: Understanding, Use, Usability, and Effectiveness of Relevance Feedback in Interactive IR**

In a doctoral dissertation project funded in part by NIST, Jürgen Koenemann (1996) (see also, Koenemann & Belkin, 1996) investigated, in an experimental setting, the effectiveness of relevance feedback in IR and the relative effectiveness of relevance feedback with different levels of user understanding and user control. Although there were several parts to this study, and although it was also connected to our TREC-4 study, only the experimental aspect will be discussed here.

The experiment asked sixty-four novice (in IR) subjects to do one different TREC search task in each of two different versions of the Rutgers InQuery (RU-INQUERY) system. In total, there were four different versions of the system: (1) without relevance feedback (the *control* system); (2) with relevance feedback with neither control nor explanation of how relevance feedback works (the *opaque* system); (3) with relevance feedback with an explanation of its workings by both an initial tutorial and by the display of what terms were added to the subject's query through relevance feedback (the *transparent* system); and (4) with both explanation of, and control over, relevance feedback through the tutorial and by presenting the terms which the system would add through relevance feedback for the user to choose (the *penetrable* system). All subjects did one search in the control system, and then subjects were randomly assigned to one of the four systems for their second search. The searches were on two different TREC topics, so all subjects searched the same topics.

Several new and important results came from this experiment. This study demonstrated, for the first time, that relevance feedback is effective in interactive IR—i.e., performance in those systems which provided rel-
Relevance feedback was significantly better than in the one which did not. Furthermore, the penetrable system, in which the subjects had explicit control over relevance feedback by choosing the terms which it would add to the query, performed consistently better than all other systems, both in terms of effectiveness (precision and recall) and in terms of effort required to reach a specific level of effectiveness. And finally, the penetrable system was consistently preferred over the control system to a greater degree than the other two relevance feedback systems. All this directly suggests not only that people can learn to use and understand relevance feedback effectively in a short period of time, but also that relevance feedback is most effective when the user can both understand and control how it works.

**TREC-5: The Understanding, Use, and Utility of Positive Relevance Feedback**

The next project in this program intended to build on the results of the TREC-4 and NIST projects described earlier. In particular, we intended to investigate, in an experimental setting, the use and effectiveness of systems with positive relevance feedback only versus those with positive and negative relevance feedback, and also to compare, in a more natural setting than in the NIST experiment, the effectiveness of user control of relevance feedback. This was all to be done in the framework of the TREC-5 Interactive Track experimental design, which had quite special characteristics (see Harman, 1997, for details).

One of the problems that the TREC Interactive Track has faced is of comparability between systems at different sites. Another is the tension in interactive IR experimental design between trying to account for topic variability (e.g., by having many topics) and trying to account for searcher variability (e.g., by having many searchers or searches). The TREC-5 Interactive Track attempted to address these issues by proposing an investigative design in which the IR task was limited to what was thought to be just one kind of IR problem (this was an attempt to control, to some extent, for topic variability). Twelve topics of this type were chosen (this was a compromise between lots of topics and lots of searchers), a new performance measure specific to this task type was developed, and an experimental design in which all searchers, at all sites, were to do the same tasks and topics in a single common control IR system, and then another set of common topics in the local experimental system (this was an attempt at comparability between systems). The specific experiment details are not too important here; it is sufficient to say that if at one site one wished to investigate the performance of one experimental system, that a single complete round of the design, which would allow some statistical inference, required twelve subjects each doing three searches in the control system and three searches in the experimental system.
For a variety of reasons, none of the participants in the TREC-5 Interactive Track was able to implement the specified investigative design. However, at Rutgers it was decided to follow the design model but without the control system which was unavailable. For a variety of reasons, we were unable to implement the systems with negative relevance feedback and user control of feedback which we had hoped to test. Therefore, we changed the foci of the study to some different issues. One was to study people's understanding of the task itself since it was unclear whether the task was realistic and understandable (it was intended to be analogous to a "keeping up to date" activity, the subjects being asked to find examples in the database of the different topics or aspects of some issue); another was whether the measure that was chosen was a good indicator of performance and whether it was a feasible measure of performance; a third was to investigate the understanding, use, and utility of positive relevance feedback for this particular task; and the fourth was to investigate the range of performance by different searchers on the same task and by the same searchers on different tasks (an attempt at investigating the issues of searcher and task variability).

The data from this study are still being analyzed; however, some preliminary results are available. From a methodological point of view, we found that our subjects had no difficulty understanding the task and relating it to some normal searching behaviors. However, the task itself was perceived as difficult to perform, requiring much interaction with, and interpretation of, documents after they had been retrieved. The measure that was developed, although it has some problems, seems suited to the task and appears to be a reasonable way to compare performance between subjects and between systems. But doing the necessary evaluation of documents with respect to this task is extremely time-intensive, and the understanding of the task by the evaluators and by the subjects is not always congruent. The experimental design itself seems to have worked, especially in terms of achieving homogenous distribution of topics in groups.

In terms of performance within our system, we found that the subjects can understand relevance feedback in this context, but that they cannot use it effectively in order to perform this task. Problems associated with control arose, as previously, but more important was the specific nature of positive relevance feedback that effectively supports people in finding more documents which are similar to the ones they like, but which does not support them in finding documents which are dissimilar—i.e., treating a different aspect of the same topic. This finding suggests that specific different IR functionalities and features are required for supporting different kinds of IR tasks. For instance, negative relevance feedback, which supports people in indicating that they do not want to see more documents like this one, could be useful in this task. Or, given the nature of the problems that subjects had with this task, features which display in
a compact and understandable manner the presence of different topics within documents might best support this type of browsing. Finally, although we had hoped that constraining the topics to a single task type would reduce variability, we found that there was quite high intra- and inter-searcher variability in performance, which we have been unable to relate to any of the other variables in the situation—e.g., subject knowledge, experience, and general difficulty of topic.

The results of the TREC-5 study have been largely suggestive rather than conclusive. The study results have led to acceptance of the general model of the TREC-5 experimental design, which we have adopted for the TREC-6 study, and they have confirmed that people can understand and use relevance feedback, and also that they have specific kinds of communication and other problems with it which could be addressed by modifying its implementation. This has led to designing the TREC-6 study explicitly to investigate user-controlled positive plus negative relevance feedback. Furthermore, these results suggest that different forms of visualization or display of databases, search results, and individual documents are necessary for supporting different kinds of IR tasks or, more generally, different kinds of interactions with information. These results are in congruence with what we have been finding in the past, and also with a model of information retrieval as interaction with information, which is being used here as the basis for a longer-term project on highly adaptive IR systems.

**TIPSTER Phase III: Understanding and Supporting Multiple Information-Seeking Strategies**

In September 1996, we began work on a three-year project within the TIPSTER program, with the eventual goal of developing IR systems that can adapt, during the course of a single information-seeking episode, to support a variety of different information-seeking behaviors (or interactions with information). The project proposal and related documents are available on the home page for our project (http://www.scils.rutgers.edu/tipster3/). This project has several theoretical and empirical bases; I would like to discuss one briefly that is strongly related to the work that has been discussed above.

There are a variety of ways to consider IR and interaction in IR. In general, the IR system has been considered as based on the fundamental processes of representation and comparison of texts and information problems, with the goal of retrieving relevant documents, and with other processes such as judgment, interaction, and modification being supportive of the comparison and representation processes. An alternative view, which I have presented earlier (Belkin, 1993, 1996), is that a person's interaction with information is the central process of IR, and that the other processes—such as comparison and representation—can be construed as tools
for supporting effective interaction. Furthermore, accepting the idea of multiple ISSs, and therefore of multiple kinds of interactions with information, and drawing upon results such as those represented by our program of research in interactive IR, leads to the belief that different kinds of interaction will be best supported by different implementations of the various IR processes. Finally, based also on our results, in particular those from the TREC-4 project, we note that persons engage in multiple ISSs in the course of a single information-seeking episode (even if the system in which they are interacting is not terribly well suited to the support of the different ISSs). Taken together, these factors lead us to conceive of IR as represented in Figure 2.

![Figure 2. Information Retrieval as Support for Interaction with Information (Belkin, 1996)](image)

The model of Figure 2 is meant to describe an information-seeking episode as it proceeds through time as a series of different kinds of interactions between the searcher and various information objects; for each kind of interaction, a different combination of specific different implementations of each of the IR processes is chosen as being the best available for support of that interaction at that time. For example, the kinds of comparison, representation, and so on techniques that will best sup-
port a person in interactions which are intended to learn about the contents of a database will not be the same as those that support the person in browsing from one related item to another, nor will these be the same as those which will support the person in searching for some documents on a topic, or for evaluating documents with respect to an information problem, and so on. In the TIPSTER Phase III project, we will use this model as the basis for the system design and for research, which follows from, and extends, the projects described above in the relationships between peoples’ goals, intentions, and situations; their interactions with information resources and information objects; the sequences of such interactions; and the combinations of IR techniques that will best support the various interactions.

GENERAL RESULTS OF THE RESEARCH PROGRAM

Although each of the projects described earlier has its own specific results, it is clear that these results also hang together in a coherent context. To summarize, we can say that the following picture of interactive IR emerges from these results:

- Having (and using) multiple representations of the information problem increases effectiveness.
- People can understand and use new system features with reasonable effectiveness, depending on the strength of their model of IR.
- Both ranked output and automatic relevance feedback are perceived as useful for certain tasks and can be used effectively.
- Systems providing query expansion through relevance feedback are more effective than those which do not.
- User understanding and control of relevance feedback leads to better performance and greater "satisfaction."
- People have a wide variety of "normal" searching strategies, many of which depend explicitly on opportunistic interaction.
- People would like more understanding and control of system features and would like some form of negative relevance feedback.
- Support for user interaction with information—i.e., incorporating the user into the information system—is effective.
- Evaluation of interactive IR is extremely difficult.

IMPLICATIONS FOR SUBJECT SEARCHING

In some ways, the results of our research program, as they refer to subject searching, are not terribly surprising, but it seems that, taken together, they have some rather important implications. For instance, it is clear from our results that searching for information (in particular, but
not exclusively through subject searching) is not only a highly complex task but also a highly interactive task. This suggests that subject searching might best be thought of as a series of interactions in which a variety of approaches are tried out, and in the course of which a variety of changes, both in the searcher and in the rest of the system which supports the searcher, take place. This may further suggest that systems to support subject searching be explicitly designed to support this kind of interaction, taking advantage of techniques like relevance feedback to support the interactions. Since it seems that each type of interaction requires a different type of support, this suggests that support for subject searching be thought of as being not one activity or process but rather some semi-structured sequence of several support features, each specific to some particular aspect of subject searching.

Although the overall goal of an information-seeking episode may be topic-related—i.e., having to do with subject searching—it is clear that interim interactions may be for quite different purposes. For instance, Cool (1997) has analyzed the thinking aloud data from the TREC-4 practice search and has identified a class of behaviors having to do with "situation assessment," and with goals needing to be met before effective topic-related interaction can take place. Some such interim goals include the establishment of authority, attunement to the information and the resource, and understanding of norms of communication in the specific situation. Again, this suggests that support for subject searching will necessarily include support for a variety of other activities, and that these will need to be considered in IR system design.

These arguments, and the results on which they are based, suggest a particular structure for IR systems—i.e., IR systems should be construed as multi-party interactions in which the user and the other components of the system cooperate and collaborate as responsible actors in the system. For instance, it appears that relevance feedback or similar techniques can be understood as conversations between the user and the intermediary about how best to represent an information problem. As far as subject searching is concerned, the implications of this design concept are that all of the parties in the system need to collaborate actively in the process of the search.

CONCLUSION

In summary, through the research program at Rutgers on interactive IR, we have established, and to some extent fleshed out, a coherent view of interactive IR that can be built on and extended progressively to address increasingly specific IR system design issues. In addition to this structure, I believe we are justified in drawing some conclusions about the nature of interactive IR itself in terms of how we might think of IR and
about the design of IR systems. These include:

- Interaction is the key to successful IR (and therefore to successful subject searching).
- Because interaction is so important, so-called "intelligent agent" models, which attempt to distance the user from the information resources, are unlikely to work well in the general subject searching (and indeed IR) case.
- Support for multiple ISSs within a single interaction will make subject searching more effective than just support for topic-related interaction(s).
- Users are actors in the IR system, as are the other components of the system, and the IR system should be designed on this basis.

REFERENCES


